# HIGH RESOLUTION STR ANALYSIS USING NEXT GENERATION SEQUENCING

## CROSS-REFERENCING

This patent application claims the benefit of provisional application Ser. Nos. 62/175,985 filed on Jun. 15, 2015, and 62/200,904 filed on Aug. 4, 2015, which applications are incorporated by reference herein.

## GOVERNMENT SUPPORT

This invention was made with Government support under contract 2013-DN-BX-K010 awarded by the United States Department of Justice. The Government has certain rights in the invention.

## BACKGROUND

Microsatellites, otherwise called STRs, have multiple alleles that are defined by variation in the number of motif unit repeats. Given their multi-allelic characteristics, they have greater heterozygosity than single nucleotide polymorphisms (SNPs). STR polymorphisms are the result of motif insertions or deletions (indels), arising from slippage errors during DNA replication or recombination events. The diversity of microsatellite alleles is attributable to STR mutation rates (10-2 events per generation) that are significantly higher than the mutation rate for SNPs which are reported to be 10-8 events per generation. Due to their multi-allelic characteristics, STR genotyping has proven useful for the genetic characterization of individual, subpopulations and populations. Moreover, genotyping with approximately 20 STRs can identify an individual with high confidence, enabling its universal application for genetic identification in forensics.

STR genotyping relies on multiplexed PCR amplification of microsatellite loci followed by analysis based on size discrimination with capillary electrophoresis (CE). Forensic genetics employs the CE-based method for nearly all cases of genetic identification. However, this approach has many limitations. First, CE genotyping assays are restricted to thirty STR amplicons or less because of the inherent challenges of multiplexing PCR reactions. Second, CE has low analytical throughput, typically in the tens of markers. Third, PCR amplification of microsatellites introduces indel artifacts, also known as "stutter", that can obscure true genotypes, particularly when alleles are close in size. Finally, current STR genotyping methods have difficulty resolving alleles in DNA mixtures that are composed of multiple individual genomes. In forensic genetic analysis, it is nearly impossible to distinguish a specific individual DNA sample amongst multiple contributors, particularly when a specific component exists at a low ratio.

Next generation sequencing (NGS) assays have been developed for the analysis of STRs. These include whole genome sequencing (WGS), targeted sequencing using bait-hybridization capture oligonucleotides and multiplexed amplicon sequencing methods that include molecular inversion probes. Regardless of the approach, current NGS methods for STR analysis have significant limitations. STRs' repetitive motifs complicate traditional alignment methods and lead to mapping errors. Sequence reads that span an entire STR locus are the most informative for accurate genotyping. However, many NGS approaches produce reads that truncate the STR sequence, resulting in ambiguous genotypes.



STR genotypes can be determined from WGS data. However, the read coverage of an intact STR locus varies greatly with the standard WGS coverage (e.g. 30× to 60×) and reduces the reads with intact microsatellites. Lower coverage translates into decreased sensitivity and specificity for detecting microsatellite genotypes. Consequently, accurate STR genotyping requires much higher sequencing coverage than is practical with WGS, particularly in cases of genetic mixtures composed of different genomic DNA samples in varying ratios.

Targeted sequencing can improve STR coverage but current methods have limitations. For example, targeting STRs with bait-hybridization enrichment requires randomly fragmented genomic DNA—this reduces the fraction of informative reads containing a complete microsatellite to less than 5%. Furthermore, enrichment for STR loci is complicated by repetitive sequences with potential off-target hybridization. Sequencing library amplification or PCR-dependent multiplexed amplicons lead to significant increase in stutter errors.

## SUMMARY

A method for analyzing short tandem repeats (STRs) is described herein. In some embodiments, the method comprises: (a) separately digesting a first portion of a genomic sample at a defined site that is upstream of an STR and a second portion of the sample at a defined site that is downstream of the STR; (b) fragmenting the cleavage products; (c) ligating adaptors to the fragmentation products; (d) selectively amplifying: part of the top strand but not the bottom strand of the ligation products derived from the first portion of the genomic sample, and part of the bottom strand but not the top strand of the ligation products derived from the second portion of the genomic sample; (e) sequencing at least some of the amplification products to produce a plurality of top strand reads and a plurality of bottom strand reads; and (f) counting the number of STR repeats in a sequence read. This count may provide an allele-specific count of the number of STR repeats at a particular locus in the genome of the individual.

In some embodiments, the sequencing step (e) is paired-end sequencing, meaning that both ends of a strand are sequenced. In these embodiments, the method comprises, prior to the counting step (f), eliminating sequence reads that do not contain the sequence of a primer used in step (d). In some embodiments, the number of STR repeats counted is validated as being accurate using a sequence read obtained from the other strand, which can be identified because it contains the sequence of the primer used in step (d).

Kits for performing the method are also provided.

## BRIEF DESCRIPTION OF THE DRAWINGS

Certain aspects of the following detailed description are best understood when read in conjunction with the accompanying drawings. It is emphasized that, according to common practice, the various features of the drawings are not to scale. On the contrary, the dimensions of the various features are arbitrarily expanded or reduced for clarity. Included in the drawings are the following figures:

FIG. **1** shows some of the principles of some embodiments of the present method.

FIG. **2** shows one implementation of a bioinformatics analysis workflow.

FIG. **3** shows how strand-specific PCR can be done on a solid support.